A System for Enhancing Human-level Performance in COVID-19 Antibody Detection

Victor Henrique Alves Ribeiro, Gabriela Steinhaus, Evair Borges Severo, José Raniery Ferreira Junior, Luiz José Lucas Barbosa, Marcelo Cossetin, Marcus Vinícius Mazega Figueredo

[1]Hilab – Curitiba, PR – Brazil

{victor.ribeiro, gabriela.steinhaus, evair.severo}@hilab.com.br

{jose.raniery, luiz.barbosa, marcelo, marcus}@hilab.com.br

Abstract. *The world currently suffers from the global COVID-19 pandemic. Billions of people have been impacted, and millions of casualties have already occurred. Therefore, it is of extreme importance to identify individuals contaminated by SARS-CoV-2, allowing governments to plan actions to reduce further impacts. In this context, this work employed machine learning to improve the detection of SARS-CoV-2 antibodies in blood exams. Models have been developed in a real-world scenario with 500 thousand exams and were deployed in a remote laboratory for experiments. Results indicate that the models averaged sensitivity and specificity of 95%, and thus, they could aid COVID-19 antibody detection and the decision-making process of biomedical specialists.*

1. Introduction

Since 2019, the world has been suffering from the Corona Virus Disease - 2019 (COVID-19) pandemic. Countries were closed in lockdown, people started to wear masks and have their body temperature measured to prevent new infections, which was not enough to avoid overcrowded Intensive Care Units (ICUs) and collapse of entire Health Systems, leaving patients without treatment [Yuan et al. 2020].

Beyond preventive measures, another important factor that enables governments and health agencies to measure the virus spread and to take action in fighting the pandemic is testing, which is highly recommended by the World Health Organization (WHO) [Beeching et al. 2020].

In this context, Hilab rises as a remote laboratory company based in Brazil, which has developed a methodology to test patients for COVID-19 Immunoglobulin G (IgG) and Immunoglobulin M (IgM) antibodies anywhere in the country and receive a result in approximately fifteen minutes, enabling an increase in the number of tests performed in a short period, with lower costs.

The company is constantly investing in technology improvements, such as Machine Learning (ML) techniques, to expand human-level performance in the exam analysis, improving the results' accuracy of the exams and reduce the time of this process [Ferreira Junior et al. 2021], as occurs in the detection of COVID-19 antibodies. Therefore, this manuscript presents a real-world application of artificial intelligence techniques to enhance human performance in analysing exams to detect antibodies against COVID-19 in blood samples.

To support the detection of COVID-19 IgG and IgM antibodies in blood samples, this work proposes the use of machine learning models to aid the decision-making process of biomedical scientists. Given the great number of exams performed by the company, Hilab makes use of a substantial dataset of professionally annotated samples to build different Multilayer Perceptrons (MLPs) to detect positive exams IgG-only, positive exams IgM-only, negative exams (no antibodies detected), and invalid exams. Moreover, the trained models are deployed into production to aid the specialists when performing new exams.

This paper is organized as follows: Section 2 explains briefly how Severe Acute Respiratory Syndrome Corona Virus 2 (SARS-CoV-2) can be identified in patients and how Hilab is contributing to spread massive testing in the Brazilian population; Section 3 proposes the application of ML models to improve the exam results and the challenges faced when dealing with substantial quantities of data. Finally, the results obtained are presented and discussed in Section 4, while the conclusions of this study are presented in Section 5.

2. Background

This section presents the background on COVID-19 and Hilab's efforts in detecting infected individuals.

2.1. Corona Virus Disease - 2019 (COVID-19)

The COVID-19 pandemic changed how we live, work, and relate to each other. Its main symptoms are dry cough, fever, and fatigue, but many others have been related to more severe infections [Huang et al. 2020]. The recommendations to reduce the viruses' spread are social distancing, usage of masks, keeping rooms well ventilated, constantly washing hands, and coughing into a bent elbow or tissue.

The most used alternative to monitor the virus advance in the population is testing, for which there are different methods. Two of them are the swab test's methodologies and the antibodies test methodologies. The latter is the scope of this work.

The swab tests are the most indicated in the days following the first symptoms, being the detection of viral Ribonucleic Acid (RNA) by molecular methods by the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test considered the gold standard [Beeching et al. 2020]. On the other hand, 7 - 14 days after the beginning of symptoms, IgG and IgM antibodies become detectable [Zhao et al. 2020] and the Serological-Based tests are the most indicated in this period.

The antibodies tests are also used for the understanding of the occurrence of infection among different populations. They also aid the identification of patients with few or no symptoms or that have not been identified in the routine disease surveillance and the proportion of the population that may be protected against future infections. They usually need few minutes to have a reaction indicating the presence or absence of antibodies in the blood sample analyzed, are easier to perform, and require less sophisticated equipment and stocking.

2.2. Hilab

Hilab is a remote laboratory company with 16-years experience in health technology, which offers 18 different types of blood exams and is specialized in Point-of-Care Testing (POCT), which are tests performed at the site of a patient that usually take only a few minutes to react and that can lead to a change in the care of the patient [ISO 22870:2016(en) 2016]. It has scanners distributed along with more than $1,000$ cities in Brazil so that 50% of the country's population lives up to 6 km from a Hilab's scanner (Figure 1).

Figure 1. Hilab's blood scanner.

In the methodology developed by the company, the blood samples are collected and processed by the local scanner (in the point-of-care). Then, the data is transmitted to the central laboratory, where expert biomedical scientists analyze it and define its results. With this methodology, Hilab can reduce the exam's cost and time for the patient to receive the result, improve efficiency, and make exams feasible in the different locations of Brazil.

To improve the performance and achieve better results, this work deploys ML techniques to aid the expert biomedical scientists. They help to improve the accuracy of the exams' analysis and assist less skilled biomedical specialists, building increasingly better ML models and increasingly better results.

3. Machine Learning for COVID-19 Antibody Detection in Blood Samples

This section presents the steps involved in building a ML system to help biomedical scientists to detect IgG and IgM antibodies in blood samples. The process is divided into the following steps: data collection, data processing, model training, model validation, and model deployment.

3.1. Data Collection

Hilab has performed over two million COVID-19 antibody exams. Notwithstanding, all exams have been labeled by biomedical specialists that have thoroughly analyzed the data. Therefore, the remote laboratory company has a large amount of available data to train and validate machine learning models.

Initial development tests have been performed with data collected from two months selected due to their volume of daily tests. Due to the vast amount of available data, hold-out validation was selected to perform model selection [Friedman et al. 2001]. In this context, data collected from October (2020) has been entirely used for model development (training and validation), while data collected from November (2020) has been used only for model evaluation.

The development dataset comprises $293,302$ samples, while the evaluation dataset is composed of $288,984$ samples. In total, the exams can have five different outtakes: i) invalid exam (due to operational error), ii) negative exam (not IgG nor IgM), iii) positive exam - IgG-only, iv) positive exam - IgM-only, or v) positive for both IgG and IgM. The ratios for each outtake in the datasets are: invalid (4.71%), negative (86.94%), positive IgG-only (2.82%), positive IgM-only (2.85%), and positive both IgG and IgM (2.67%), which characterizes an imbalanced learning problem [He and Ma 2013].

3.2. Data Processing

Unfortunately, due to legal constraints and trademark reasons, raw data and data processing techniques are not entirely presented in this manuscript. Nevertheless, the final processed data is composed of a total of 450 numerical features.

3.3. Model Training

To train a machine learning model to classify the different classes, a binary classification structure has been selected using MLPs [Goodfellow et al. 2016]. The architecture of the MLPs uses a single hidden layer with $n = 450$ neurons in both the input and hidden layers (Figure 2). Moreover, the activation functions for the hidden and output layer are Rectified Linear Unit (ReLU) and sigmoid, respectively (Table 1), while the network is optimized by the Adam optimizer [Kingma and Ba 2014] using binary cross entropy (Equation 1) with learning rate $l = 0.001$.

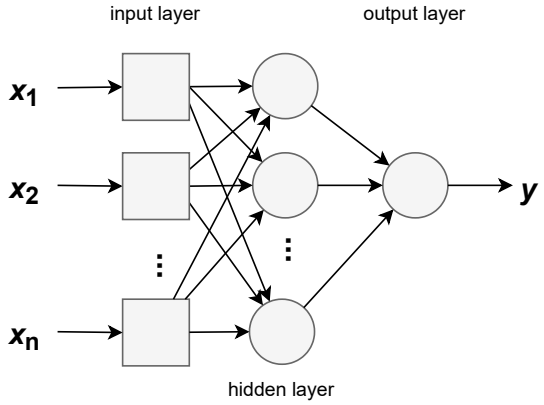

Figure 2. Binary Classification Multilayer Perceptron.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i.log(p(y_i)) + (1 - y_i).log(1 - p(y_i)) \quad (1)$$

Table 1. Architecture of the Multilayer Perceptron.

Layer	Neurons	Activation Function
Input	450	-
Hidden	450	ReLU (Equation 2)
Output	1	sigmoid (Equation 3)

$$\text{ReLU}(x) = \max(0, x) \tag{2}$$

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

The final solution is composed of four one-versus-all models. The first model focuses on detecting invalid exams, the second on negative exams, and the other two on the IgG and IgM antibodies, respectively. All models have been trained for 200 epochs with a batch size of 32 samples. The models have been trained and validated with 80% and 20% of the development set, respectively, where a patience of 30 epochs has been set to enable early stopping. In addition to this, given the imbalance ratio of the classes, the weights of the minority classes have been computed given the inverse of their respective ratios.

3.4. Model Evaluation

Once the models are all trained, the evaluation is performed on the separate test dataset given the sensitivity (Equation 4) and specificity (Equation 5), where True Positives (TP) indicates the number of condition positives correctly predicted as such, True Negatives (TN) indicates the number of condition negatives correctly predicted as such, False Positives (FP) indicates the condition negatives incorrectly predicted as positives (type I error), and False Negatives (FN) indicates the condition positives incorrectly predicted as negatives (type II error).

$$Sensitivity = TP/(TP + FN) \tag{4}$$

$$Specificity = TN/(TN + FP) \tag{5}$$

One additional criterion for the models is the ratio of predictions with high confidence. In such a sense, only predictions with high confidence are selected to assist the biomedical scientists in the decision-making process. To this end, a simple rule implies that only predictions with over 95% confidence (from the sigmoid output) are considered as high confidence predictions.

3.5. Model Deploy

Finally, once the models are trained and validated, the models are deployed on a production server. In such a scenario, the software used by the biomedical specialists sends the data to the models and returns their outputs to give additional information to the operators. Figure 3 shows part of the software's Graphical User Interface (GUI), which shows the model's outputs and includes fields for the operator to correct it. Moreover, information

of the processed data is shown to the users (only part of it is being presented due to legal constraints).

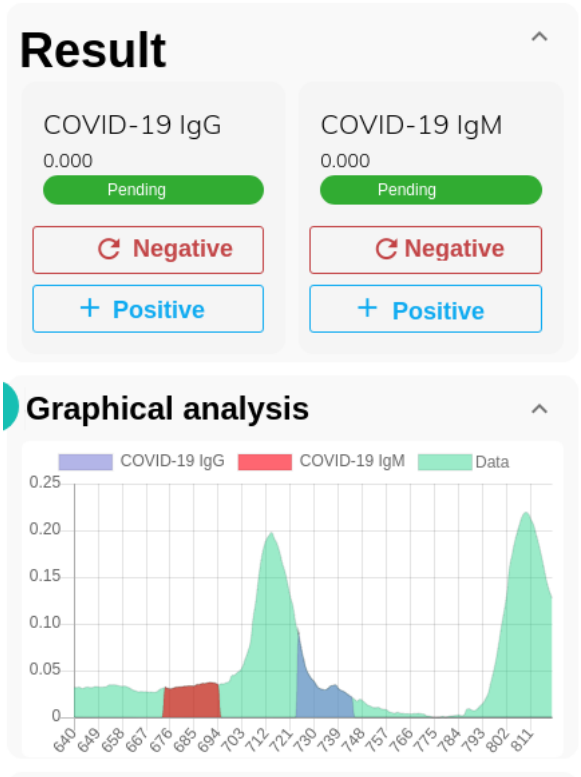

Figure 3. Models' predictions used to assist the biomedical scientists in the operator's software.

4. Results and Discussion

The evaluation of the proposed approach resulted in high scores for the classification task. Figures 4 to 7 plot the normalized confusion matrix for each of the four models on the test data set. Moreover, Figure 8 shows a dashboard that analyzes the models' performances in real-time.

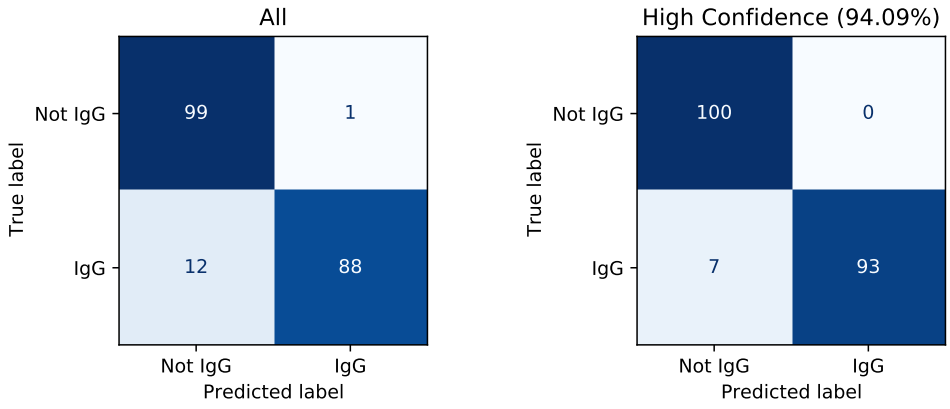

Figure 4. Confusion matrices for the IgG detection model.

The IgG model achieves high specificity (99%) and relatively high sensitivity (88%). However, such a model presents a high rate of high confidence prediction (94%), where the specificity and sensitivity increase to 100% and 93%, respectively.

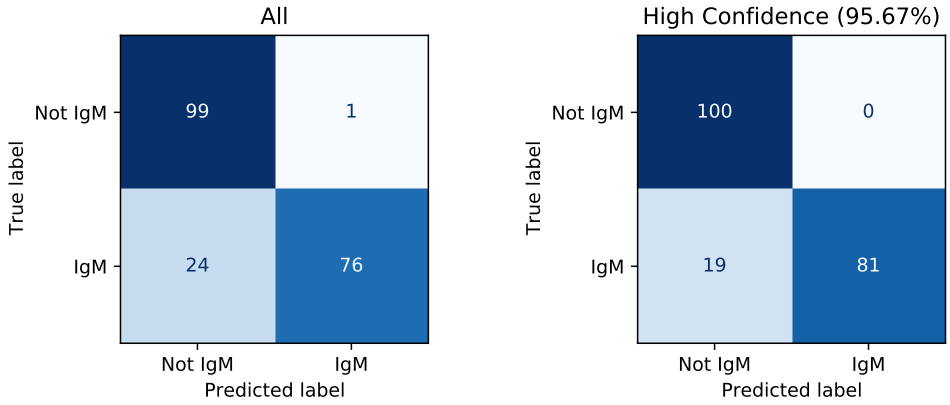

Figure 5. Confusion matrices for the IgM detection model.

The IgM model presents a more difficult task in terms of sensitivity. While the specificity for all data and the high confidence predictions achieve 99% and 100%, respectively, the sensitivities are lower. The sensitivity from all predictions to the high confidence predictions grows from 76% to 81%. Despite this, over 95% of the predictions achieve high confidence.

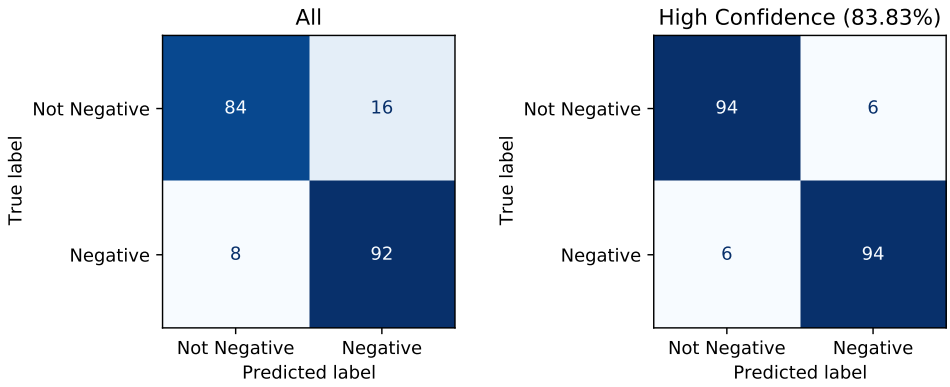

Figure 6. Confusion matrices for the negative detection model.

The detection of negative exams presents better results for true positive rate, which is caused by the imbalance ratio of the dataset. While the sensitivity and specificity for all predictions are 92% and 84%, respectively, filtering the high confidence predictions (83%) increases both metrics to 94%.

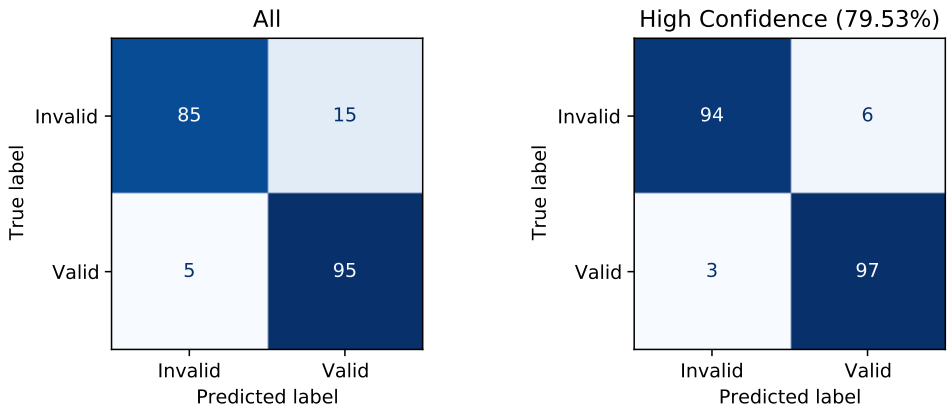

Figure 7. Confusion matrices for the invalid exam detection model.

Finally, the invalid model shows a more difficult problem to separate. While the sensitivity and specificity for all data are 95% and 85%, respectively, only 79% of the predictions achieve high confidence. By analyzing only the high confidence predictions, the sensitivity increases to 97% while the specificity increases to 94%.

Such results show a difference between the sensitivity and specificity of the models. That is mostly caused by the complexity of the problem's data, such as labeling and noise. Therefore, since the negative class composes the majority of the samples, its sensitivity achieves high results. On the other hand, since all other classes compose the minority of the samples, the specificity achieves better results. In regards to classification difficulty, the IgM model presents the lowest scores. That occurs because the raw data and the features generated for such a model present a lower signal-to-noise ratio, which is inherent to the physical properties of the exam. Nevertheless, it is important to discuss the high confidence ratio of the predictions, which is lower for the negative and invalid detection model. For the invalid model, there exists noise in the labeling, which is caused by the different biomedical specialists. While one expert may consider a sample invalid, other may consider the same sample as valid. Moreover, the negative class has a lower ratio of high confidence predictions because there is also a higher noise between such a class and the invalid one, which does not occur much for the IgM and IgG classes.

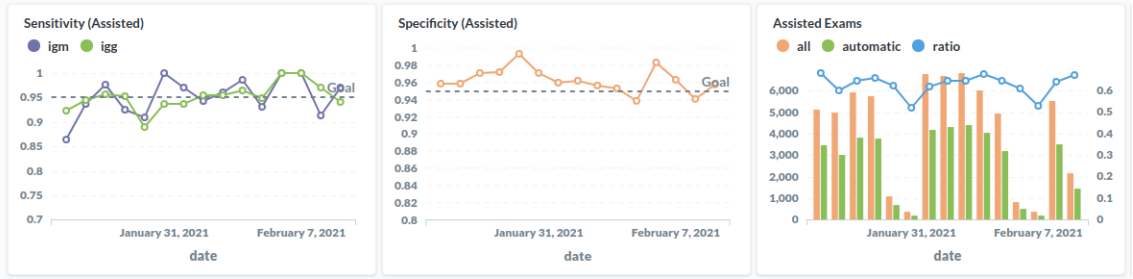

Figure 8. Crop of the dashboard for the real-time analysis of the models' performances on the deployed scenario.

Finally, Figure 8 shows part of the dashboard with the real-time analysis of the classification models in the deployed scenario. There is a goal to maintain both the sensitivity and the specificity of the system at a minimum of 95%. This result is important because it allows the biomedical specialist to analyze the exam in a shorter time and to be more confident of the result [Ferreira and Cardenas 2021], since this short time interval between taking the test and the result is crucial for decision-making and the virus' spread containing. Currently, the system can achieve such results on approximately 65% of the exams, on average. However, the greatest task is to produce more high confidence predictions and to increase such ratio. It is important to study different machine learning models and, most importantly, new feature extraction techniques to produce more representative data to train the classifiers.

5. Conclusions

This work presents a real-world application of machine learning to aid COVID-19 antibody detection in blood samples. Hilab is a remote laboratory company dedicated to improve the access to health solutions in Brazil. Using a blood scanner to perform POCT in more than one thousand locations in the country, faster and cheaper blood exams are performed and analyzed by biomedical specialists, including the detection of IgG and IgM antibodies from COVID-19.

In the pandemic context, the implementation of artificial intelligence to support COVID-19 antibodies' exam analysis is of fundamental importance, in the first place, to speed up sick patients detection and decision-making in order to reduce the virus propagation, and also to improve epidemiological studies to understand the SARS-CoV-2 spreading among a population and define new means to stop it.

Results indicate that the models achieve, on average, over 95% sensitivity and specificity in the real-world scenario, assisting the specialists in more approximately 65% of the performed exams. The improvement of the models to increase such results certainly reduce the doubts of biomedical scientists, increasing their performance by decreasing the time to return a result to the patient, lowering human mistakes, and improving the training of new specialists to perform the exams' analysis.

Nevertheless, the deployment of the models in the production server can improve the quality of the annotated data, enabling improvements in the model performance and vice-versa. Additionally, the models have been trained using default hyperparameters. Therefore, future work focuses on retraining the same models using better-annotated data and applying hyperparameter optimization. Moreover, different machine learning techniques and feature extraction techniques can be tested to improve such results. Most specifically, given the number of available data, deep learning techniques can be used in future research.

References

Beeching, N. J., Fletcher, T. E., and Beadsworth, M. B. J. (2020). Covid-19: testing times. *BMJ*, 369.

Ferreira, J. R. and Cardenas, D. A. C. (2021). The potential role of radiogenomics in precision medicine for COVID-19. *Journal of Thoracic Imaging*. DOI:10.1097/RTI.0000000000000586.

Ferreira Junior, J. R., Cardenas, D. A. C., Moreno, R. A., Rebelo, M. F. S., Krieger, J. E., and Gutierrez, M. A. (2021). Novel chest radiographic biomarkers for COVID-19 using radiomic features associated with diagnostics and outcomes. *Journal of Digital Imaging*. DOI:10.1007/s10278-021-00421-w.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

He, H. and Ma, Y. (2013). *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223):497–506.

ISO 22870:2016(en) (2016). Point-of-care testing (POCT) — Requirements for quality and competence. Standard, International Organization for Standardization, Geneva.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yuan, M., Yin, W., Tao, Z., Tan, W., and Hu, Y. (2020). Association of radiologic findings with mortality of patients infected with 2019 novel coronavirus in Wuhan, China. *PloS ONE*, 15(3):e0230548.

Zhao, J., Yuan, Q., Wang, H., Liu, W., Liao, X., Su, Y., Wang, X., Yuan, J., Li, T., Li, J., Qian, S., Hong, C., Wang, F., Liu, Y., Wang, Z., He, Q., Li, Z., He, B., Zhang, T., Fu, Y., Ge, S., Liu, L., Zhang, J., Xia, N., and Zhang, Z. (2020). Antibody Responses to SARS-CoV-2 in Patients With Novel Coronavirus Disease 2019. *Clinical Infectious Diseases*, 71(16):2027–2034.